NP-Scout: Machine Learning Models for the Identification and Visualization of the Natural Product-Likeness of Small Molecules

Ya Chen [1], Conrad Stock [1], Steffen Hirte [1] and Johannes Kirchmair [1,2,3]

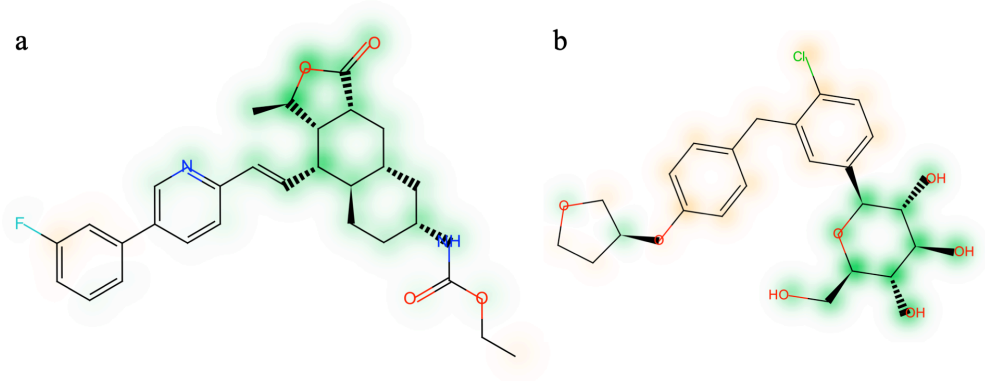
[1] *Center for Bioinformatics (ZBH), Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany*

[2] *Department of Chemistry, University of Bergen, 5020 Bergen, Norway*

[3] *Computational Biology Unit (CBU), Department of Informatics, University of Bergen, 5020 Bergen, Norway*

Natural products (NPs) remain the most productive source of inspiration in drug discovery [1]. NPs are highly diverse in chemical structure, exhibit a broad range of activities relevant to human health, and cover regions of the chemical space not or only rarely populated by synthetic molecules (SMs) [2, 3]. However, only an estimated 10% of all known NPs are readily obtainable for testing [4], and despite the rarity of materials, vendors most commonly offer mixed libraries of unlabeled (semi-) SMs and NPs.

We have devised a machine learning approach ("NP-Scout") [5] that not only allows the discrimination of NPs and SMs but also enables the quantification of NP-likeness and the visualization of regions in molecules that are NP-like. NP-Scout uses random forest classifiers trained on more than 265,000 NPs and SMs compiled from a total of 27 data sources. The suitability of two-dimensional molecular descriptors, MACCS keys and Morgan2 fingerprints was explored for model building. On independent test sets, NP-Scout models reached areas under the receiver operating characteristic curves (AUCs) of up to 0.997 and Matthews correlation coefficients (MCCs) of 0.954 and higher. The method was successfully tested also on data from the Dictionary of Natural Products, ChEMBL and further resources. Similarity maps generated from the models highlight atoms that contribute to the classification of small molecules as NPs or SMs. This allows, for example, also the identification of NP derivatives, as shown below by the example of the NP derivatives (a) vorapaxar and (b) empagliflozin (atoms highlighted in green are recognized as NP-like, atoms highlighted in orange as SM-like). The best-performing models are accessible via a free web service at http://npscout.zbh.uni-hamburg.de/npscout.



[1] G. M. Cragg, D. J. Newman, *J. Macromol. Sci. Part A Pure Appl. Chem.* **2005**, 77, 7–24.
[2] P. Ertl, A, Schuffenhauer, *Prog. Drug Res.* **2008**, 66, 219-235.
[3] Y. Chen, M.G. de Lomana, N.-O. Friedrich, J. Kirchmair, *J. Chem. Inf. Model,* **2018**, 58, 1518-1532.
[4] Y. Chen, C. de Bruyn Kops, J. Kirchmair, *J. Chem. Inf. Model,* **2017**, 57, 2099-2111.
[5] Y. Chen, C. Stork, S. Hirte, J. Kirchmair, *Biomolecules,* **2019**, 9, 43.